Online Infrared UAV Target Tracking With Enhanced Context-Awareness and Pixel-Wise Attention Modulation

Houzhang Fang, *Member, IEEE*, Chenxing Wu, Xiaolin Wang, Fan Zhou, Yi Chang, *Member, IEEE*, and Luxin Yan, *Member, IEEE*

***Abstract*— Unmanned aerial vehicles (UAVs) have been popular in many commercial and industrial applications, but they also pose great threats to urban safety and aerial security. Intelligent UAV surveillance based on thermal infrared (TIR) imaging has attracted increasing attention for its long-range monitoring ability in both day and night scenarios. However, weak UAV target features, dynamically changing UAV states, and complex background interferences present serious challenges to the accurate tracking of UAVs. To tackle these problems, we propose a novel online multiscale infrared UAV target (IRUT) tracking network (SiamCAP) incorporating enhanced context feature awareness and pixel-wise attention modulation. We first introduce a novel contrast-enhanced multiscale online re-parameterization block (CMORB) to effectively extract contrast difference intensity information between the target and the background, and transform it into a single branch for both training and inference without introducing computational overhead. Then, we construct a feature fusion modulation module (FFMM) to guide cross-layer feature aggregation. It uses low-level attention to highlight the UAV target feature in the deep layer with the novel full spatial resolution channel attention (FSRCA), which calculates pixel-wise importance without dimensionality compression. Finally, we propose a cross-attention-based updatable feature interaction module (CUFIM) to model the correlation between online updating multitemplate and search frame, which improves the model's robustness to changes in the state of UAVs and complex backgrounds. Extensive experiments on real infrared UAV datasets demonstrate that the proposed approach outperforms the state-of-the-art (SOTA) target trackers under complex backgrounds while achieving a real-time tracking speed.**

***Index Terms*— Attention mechanism, context awareness, infrared target tracking, intelligent unmanned aerial vehicle (UAV) surveillance, online re-parameterization.**

Manuscript received 3 January 2024; revised 26 February 2024 and 15 June 2024; accepted 17 July 2024. Date of publication 22 July 2024; date of current version 2 August 2024. This work was supported in part by the Open Research Fund of the National Key Laboratory of Multispectral Information Intelligent Processing Technology under Grant 61421132301 and in part by the National Natural Science Foundation of China under Grant 61971460 and Grant 62101294. *(Houzhang Fang, Chenxing Wu, and Xiaolin Wang are co-first authors.) (Corresponding authors: Houzhang Fang; Fan Zhou.)*

Houzhang Fang, Chenxing Wu, and Xiaolin Wang are with the School of Computer Science and Technology, Xidian University, Xi'an 710126, China (e-mail: houzhangfang@xidian.edu.cn; wcx1111111029@163.com; wxl@stu.xidian.edu.cn).

Fan Zhou is with Xi'an Modern Control Technology Research Institute, Xi'an 710065, China (e-mail: 18602969817@163.com).

Yi Chang and Luxin Yan are with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: yichang@hust.edu.cn; yanluxin@hust.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TGRS.2024.3432108, provided by the authors.

Digital Object Identifier 10.1109/TGRS.2024.3432108

I. Introduction

RECENTLY, unmanned aerial vehicles (UAVs) have been widely adopted in many fields, such as aerial photography, environmental monitoring, and precision agriculture [3]. However, the pervasive use of UAVs can also potentially pose great threats to both public security and aerial safety. Thermal infrared (TIR) imaging technology enables monitoring UAVs at a long distance in both day and night conditions. Accordingly, the anti-UAV technologies based on TIR imaging have attracted growing attention. Infrared target tracking is a key perception technology for UAV surveillance in the anti-UAV system, which requires tracking UAVs accurately and efficiently. However, there exist four major challenges in the task of infrared UAV target (IRUT) tracking.

1) *Weak Target Features:* In infrared UAV surveillance, the UAV targets often occupy only a small part of the entire image and are susceptible to background occlusion (OC), making it difficult to extract discriminative features.
2) *Frequent UAV Disappearance and Reappearance:* The fast movement of the UAV may cause it to disappear from the field of view of the infrared imaging platform within a certain period. Hence, the tracker needs to have the ability to recapture the target when it reappears.
3) *Multiscale Variation of UAV Targets:* The dynamic changes of UAVs in motion lead to multiscale characteristics in the observed UAV images. Therefore, the tracking method needs to be capable of detecting multiscale UAV targets, especially small-scale ones.
4) *Complex Background Interference:* UAV targets usually suffer from dynamic complex background interference during the movement, such as thick and bright clouds, trees, and bushes. The targets can be easily interfered with or submerged by these backgrounds, exacerbating the difficulty of UAV target tracking.

With the development of deep learning, many CNN-based target tracking methods have been developed [4], [5], [6], [7], [8], [9]. However, these methods fail to exploit the long-range interdependency between the template and search frame with simple linear matching. Recently, the transformer has achieved significant success in the target tracking task [10], [11], [12], benefiting from the ability to model the global relationship. However, these trackers only search for the

1558-0644 © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 05,2024 at 01:19:02 UTC from IEEE Xplore. Restrictions apply.

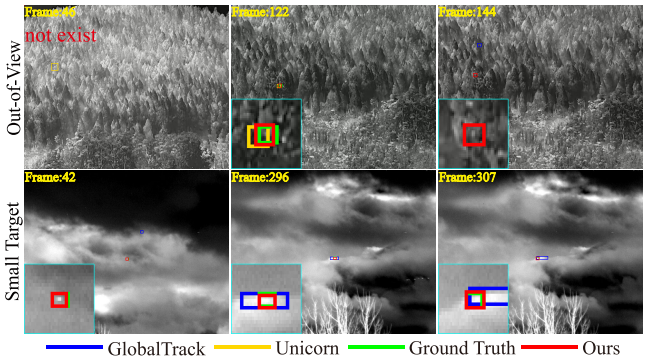


Fig. 1. Qualitative comparison of our method with the baseline long-term trackers GlobalTrack [1] and Unicorn [2] on two challenging video sequences (OV and small target). The proposed Siamese tracker achieves favorable performance for IRUTs under different complex backgrounds thanks to effective integration of the contrast-enhanced contextual features, PCAM, and online updating multitemplate. "not exist" means the target is not in the field of view.

target within a local region, which means they cannot address IRUT tracking failure due to target disappearance or fast motion (FM). To tackle these challenges, it is essential to incorporate a tracker with global re-detection capability in the anti-UAV task.

SPLT [13] and LTMU [14] jointed a local tracker and global re-detection mechanism to deal with the problem of tracking failure. The advantage of this strategy is that it can exploit the target's temporal information. However, it also brings the risk that the global tracker will not be activated when the local tracker mistracks the distractor with high confidence. Furthermore, in anti-UAV tasks, camera move and FM of UAVs are common occurrences. The frequent switching between the local tracker and global tracker leads to significant time costs, thereby restricting their practical application in real-world scenarios. In contrast to the local–global tracker, GlobalTrack [1], Siam R-CNN [15], and Unicorn [2] globally searched targets invoking target detection on the whole image. A significant benefit of this strategy is its ability to prevent error accumulation. However, these trackers can be susceptible to distractors since they only leverage the single fixed template. Robustness and real-time are crucial factors for successfully applying trackers in anti-UAV systems. In this work, we adopt a global search-based method to conduct the IRUT tracking task. Meanwhile, we enhance the model's anti-interference capability by incorporating reliable temporal information.

Although the above trackers achieve impressive performance for general long-term target tracking tasks, they can hardly deal with challenges met in anti-UAV tasks. To achieve robust tracking of IRUT, Huang et al. [16] designed a local–global tracker combined with spatial–temporal constraints to eliminate interference from similarities. Zhao et al. [17] introduced a camera motion estimation module to correct the target search range of the current frame and a model updater to reliably switch between the local tracker and the global tracker. Jiang et al. [18] leveraged the class-level and instance-level semantic information across video sequences to improve the UAV tracking performance. However, these trackers still encounter difficulties when it comes to tracking small IRUTs. First, it is challenging to fully capture the intrinsic features of small targets only using general backbone networks. Second, they often overlook the effective utilization of semantic features in deep layers which are conducive to the recognition of IRUT. Third, these trackers are less discriminative when the appearance of the target changes, as they rely solely on the initial template to match with the search frame. To improve the feature representation ability of IRUT, Wu et al. [19] proposed channel-wise attention to enhance the feature representation ability and enlarge the search region when tracking failure occurs. Fang et al. [20] proposed a Siamese network-based anti-distractor global tracker to localize multiscale IRUTs using the attention mechanism and spatial–temporal information. The attention mechanisms employed by these methods are either compressed in the spatial dimension or compressed in the channel dimension. However, this compression process inevitably leads to information loss, particularly for small targets.

Based on the above considerations, we summarize the shortcomings of existing methods.

1) Limited capacity in extracting discriminative contextual target features. The current single-branch structure at each hierarchical level of the backbone struggles to learn the distinctive representation of IRUTs.
2) Less emphasis on suitable feature fusion methods based on contextual information guidance with the uncompressed attention mechanism for multiscale target tracking. The existing feature fusion guided by the compressed attention mechanism is not effective due to weak target features and complex background interference.
3) Little attention to the online template update mechanism. The existing single template-based global search methods are susceptible to interferences from similar targets, leading to frequent tracking drifts.

To deal with the above issues, in this article, we integrate enhanced feature awareness and pixel-wise attention modulation into the Siamese network for online multiscale IRUT tracking. To be specific, we first devise a novel contrast-enhanced multi-scale online reparameterization block (CMORB) in the training stage to subtly strengthen UAV target features based on the difference between target and background as well as capture local contextual feature information at each hierarchical level of the backbone. The multibranch can be converted to single-branch in both training and inference phases without introducing additional computational burden.

We then design a feature fusion modulation module (FFMM) to guide the cross-layer information exchange of the low-level and high-level features by integrating a novel full spatial resolution channel attention (FSRCA) mechanism. The FSRCA mechanism can adaptively strengthen the UAV target features pixel-wise using channel attention without spatial and channel compression, leading to significant performance improvements, especially for small-scale IRUTs. Furthermore, we adopt the pixel-wise cross-channel interaction strategy to reduce the computational complexity of FSRCA, which can be efficiently implemented via depth-wise convolution.

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 05,2024 at 01:19:02 UTC from IEEE Xplore. Restrictions apply.

To further boost the long-term tracking performance of IRUTs, we introduce the cross-attention-based updatable feature interaction module (CUFIM), which adaptively models the interdependency between multiple templates and search frames, improving the model's awareness of UAV targets in search frames. Meanwhile, the incorporation of multiple templates' online updating mechanisms during tracking allows the model to perceive changes in the UAV target's state and the complicated background. Based on CUFIM, we construct three global tracking branches to achieve the target recapture when it reappears. The qualitative comparison of our method with the other two baseline long-term trackers is shown in Fig. 1.

The main contributions of this article are summarized as follows.

1) We propose a novel online long-term target tracking network (SiamCAP) of infrared UAVs, which can consistently and globally track multiscale IRUTs under complex backgrounds while achieving real-time inference speed.
2) We present a novel CMORB to boost the discriminative contextual feature representation capability at each hierarchical level, which can improve the tracking performance of IRUTs without introducing extra computational costs.
3) We introduce a novel FSRCA mechanism and integrate it before each global search branch to highlight the UAV target features pixel-wise in the spatial and channel dimensions, which can further guide the refinement of the high-level features with low-level features in a bottom-up manner.
4) We construct a new CUFIM that effectively captures the temporal association information between the multiple templates and search frame for the model to be aware of more recent target state changes.

The remainder of this article is arranged as follows. In Section II, we present a brief review of the relevant studies on target tracking. In Section III, we provide the architecture of the proposed network in detail. In Section IV, we show extensive results to evaluate the proposed method. Finally, we conclude the proposed method in Section V.

II. Related Work

This section concisely reviews the existing Siamese network-based tracking methods, covering the CNN-based and transformer-based methods. Furthermore, long-term target tracking, multiscale feature fusion, and infrared target tracking methods are also explored.

A. Siamese Network-Based Target Tracking

Recently, Siamese network-based target tracking methods have been receiving significant attention because of their equilibrium of accuracy and efficiency. Bertinetto et al. [4] proposed a fully convolutional network SiamFC by modeling target tracking as a cross correlation problem between the template and search frames. Li et al. [6] formulated the tracking task as a one-shot detection by introducing the region proposal network (RPN) into the Siamese framework. Li et al. [7] further integrated deep network ResNet into the Siamese trackers to improve the feature representation ability. Guo et al. [21] proposed an anchor-free tracking framework that significantly reduces the number of hyperparameters that need to be tuned. However, these methods utilize correlation operation in a local region without using global information aggregation, making the model susceptible to interference. In recent years, thanks to its capability of capturing long-range dependency, Transformer has been successfully applied to tracking tasks. Ye et al. [10] leveraged joint feature extraction and relationship modeling to enhance the model's discriminative power. Gao et al. [11] used the estimated intersection-over-union (IoU) score to evaluate the prediction quality and introduced a two-branch cross-attention design combined with long-term and short-term references. Gao et al. [12] utilized a token division module to split the search frame token into two parts and only an appropriate subset of search tokens was selected to interact with template tokens. However, these trackers are inability to re-detect targets once they have temporarily disappeared from the field of view and reappeared subsequently. In this work, we propose a global search paradigm-based tracker combined with a new CUFIM, which can effectively handle the challenges posed by the occurrence of target disappearance and reappearance and changes in target appearance.

B. Long-Term Target Tracking

The current long-term tracking framework can be divided into two categories: one is equipped with a local tracker with the global re-detection mechanism; the other one is to globally search the target within the input image. Yan et al. [13], Dai et al. [14], Wang et al. [22], and Xu et al. [23] combined a local tracker with a global tracker to deal with tracking failure situations. Yan et al. [13] proposed to generate candidate target proposals based on SiamRPN, and searched targets locally or globally by the skimming and perusal modules. Dai et al. [14] trained a meta-updater with a series of long short-term memory (LSTM) modules to reliably decide the switch between the local tracker and global re-detector. In contrast to the local–global tracker, Yan et al. [2], Fang et al. [3], and Voigtlaender et al. [15] adopted a global search strategy to track the target in each frame. Voigtlaender et al. [15] and Huang et al. [1] constructed global instance search-based trackers by drawing ideas from two-stage target detectors. However, these trackers are developed only for visible light data with rich details. Their effectiveness in IRUT tracking tasks remains suboptimal because their network architectures lack the ability to effectively model small IRUT features. In this study, we leverage the multibranch architecture combined with contrast-enhanced design and the full-resolution channel attention mechanism to construct an enhanced representation network for effectively extracting small-scale UAV target features.

C. Multiscale Feature Fusion

Existing research [6], [21], [24], [25] shows that incorporating feature fusion between adjacent layers into tracking

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 05,2024 at 01:19:02 UTC from IEEE Xplore. Restrictions apply.

frameworks can improve multiscale target tracking. However, when it comes to IRUTs, which are typically smaller compared to targets in general tracking tasks, these works simply concatenate or add the features of the adjacent layers making it difficult to effectively utilize the deep layer features (especially for small-scale ones).

Some works [26], [27], [28] try to introduce attention mechanisms to guide the information exchange between deep and shallow layer features. Li et al. [29] proposed to incorporate global channel attention [30] for guiding the feature fusion between the shallow layer and deep layer. This global channel attention can effectively highlight features for large-scale targets, which are distributed more globally. To enhance small-scale target features, Dai et al. [26] proposed point-wise channel attention to modulate the high-level features with low-level ones. However, the application of channel compression in [26] still hinders it from accurately representing small-scale target features. In this work, we construct a new FSRCA without spatial and channel compression that is crucial in boosting and preserving subtle small-scale UAV target features pixel-wise.

D. Infrared Target Tracking Methods

In recent years, infrared target tracking has drawn growing attention. Liu et al. [31] proposed integrating global semantic features and local structural features into a deep multilevel similarity model of the Siamese framework to strengthen the discrimination ability of the model. Fang et al. [20] proposed a real-time anti-distractor infrared UAV tracker, which can globally track multiscale UAV targets based on the idea of target detection and discard distractors by leveraging spatial–temporal information. Jiang et al. [18] constructed the anti-UAV datasets and adopted the class-level semantic modulation and instance-level semantic modulation to obtain robust and discriminative features between UAV targets and similar distractors. However, IRUTs are small in scale and change in pose during motion, making it difficult to effectively extract the discriminative features. Although these methods can track infrared targets, they pay little attention to modeling the contextual information of the targets subtly. In this study, we explore the representation of context feature information and incorporate it into the target tracking network to boost the performance of IRUT tracking.

III. Methodology

Consistent IRUT tracking is a challenging task in complex and dynamic scenarios. In this work, we propose a real-time tracking architecture based on a multibranch feature extractor with two contrast-enhanced branches, a contextual modulation module with pixel-wise FSRCA, and a temporal contextual information awareness feature interaction module. The proposed architecture is shown in Fig. 2.

A. Contrast-Enhanced Multiscale Online Re-Parameterization Block

The IRUT has the characteristics of weak features, lack of texture information, and dynamic multiscale variation. A robust feature extraction backbone network is essential for effectively extracting IRUT features. In this work, we construct a CMORB to capture both local contrast difference intensity between target and background and rich local contextual features, thus effectively improving the model's ability to learn discriminative IRUT features. Infrared imaging typically leads to UAVs exhibiting a higher temperature than the background, resulting in a contrast between the target and the background. However, merely using normal convolutional operations is insufficient for capturing this difference information. We additionally introduce two contrast-enhanced branches into the base block of the backbone network which can strengthen the module's capability to extract the difference information. Furthermore, utilizing the convolution kernels of different sizes can boost the richness of multireceptive field features which facilitates the multiscale feature extraction capability.

Although multibranch convolutional operations can improve tracking performance, they significantly increase the computational cost. To improve the tracking performance while saving the training cost and maintaining high efficiency, we propose to introduce the online convolutional re-parameterization strategy [32] into our block design. This technique leverages a linear scaling layer realizing the merge of multiple convolution branches into a single branch during training and inference phases, which not only greatly shortens the training time but also avoids introducing additional inference time consumption.

In this research, we employ the ConvNeXt-Tiny [33] as our baseline backbone network. However, the original ConvNeXt model only uses a 7×7 depth-wise convolution in each basic block, thus its limited ability to effectively extract fine-grained information from IRUT. For greater feature extraction ability of the network for IRUTs, we additionally introduce cross central difference convolutions [34] in each block to aggregate difference information between the center pixel and surround sparse neighbors in two directions and modify the structure to the separable depth-wise convolutions, i.e., the 3×3 horizontal/vertical (HV) cross central difference depth-wise convolution (C-CDDC) path and 3×3 diagonal (DG) C-CDDC path. Compared with normal convolution, the center difference convolution pays more attention to the difference between the target and the surrounding background, while the normal convolution pays more attention to local details and semantic features. However, when the target is immersed in the background, such as in the presence of heavy clouds, it can be challenging to extract the contrast difference feature. Thanks to the multibranch structure, the model benefits from the aggregation of information extracted by various convolutional operations, resulting in an improved feature extraction ability in different complex scenarios. In addition, we replace layer normalization (LN) after each convolution with a linear scaling layer for online re-parameterization. The CMORB is illustrated in Fig. 3. During both the training and inference phases, the multiple paths following the linear scaling operation are merged into a single convolution path, and then they go through two point-wise convolutions for channel interaction. Consequently, CMORB is able to effectively exploit IRUT

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 05,2024 at 01:19:02 UTC from IEEE Xplore. Restrictions apply.

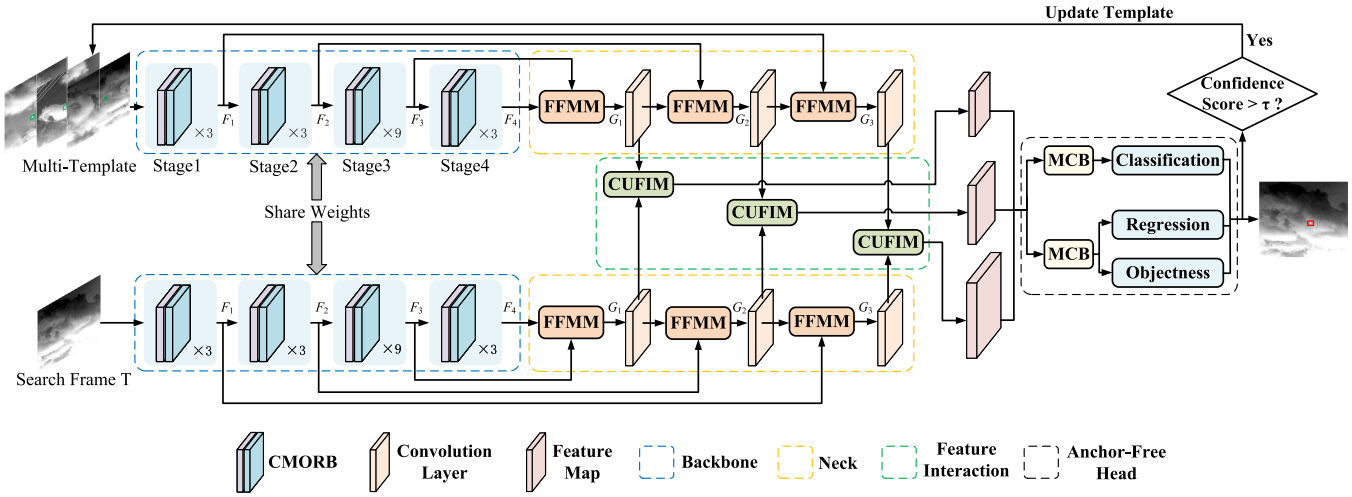


Fig. 2. Overview of the proposed architecture. The four main components of the framework are marked by dotted boxes of different colors. The backbone consists of four stages based on the CMORB, and the multibranch is fused into a single convolution branch in the training and inference phases for reducing training cost and maintaining high inference speed. The FFMM exploits the pixel-level FSRCA to guide contextual information exchange of the high-level features with the low-level features. The CUFIM enhances the model's identified capabilities for UAV targets under the complex background by incorporating multiple historical templates. MCB denotes multiconvolution blocks. The extra introduced templates will be updated when the confidence score, which is obtained by multiplying the classification score and objectness score, is higher than the threshold τ.

features, capture discriminative multiscale UAV features, and ensure high inference speed in the inference phase.

We now describe how to merge multiple convolution paths into a single 7×7 depth-wise convolution path for efficient training and inference. The C-CDDC can be formulated as

$$\begin{aligned} y(p_0) &= \theta \cdot \underbrace{\sum_{p_n \in \mathcal{R}} w(p_n) \cdot (f(p_0 + p_n) - f(p_0))}_{\text{central difference convolution}} \\ &\quad + (1-\theta) \cdot \underbrace{\sum_{p_n \in \mathcal{R}} w(p_n) \cdot f(p_0 + p_n)}_{\text{vanilla convolution}}, \\ &= \sum_{p_n \in \mathcal{R}} (w(p_n) - \theta \cdot g(p_n)) \cdot f(p_0 + p_n)^{1} \end{aligned} \tag{1}$$

$$g(p_m) = \begin{cases} \sum_{p_n \in \mathcal{R}} w(p_n), & \text{if } m = 0 \\ 0, & \text{elsewise} \end{cases} \tag{2}$$

where y, w, f, and p_0 denote the output feature map, the weights of the convolution kernel, the pixel value in the feature map, and the current location on both input and output, respectively. Both p_n and p_m represent the neighboring locations around p_0, and $p_n, p_m \in \mathcal{R}$, $p_0 = (0, 0)$. Specifically, $\mathcal{R} = \{(-1, 0), (0, -1), (0, 0), (0, 1), (1, 0)\}$ in HV direction and $\mathcal{R} = \{(-1, -1), (-1, 1), (0, 0), (1, -1), (1, 1)\}$ in DG direction. The hyperparameter θ balances the contribution between contrast intensity information and local contextual features. Formally, $\mathbf{D}_i \in \mathbb{R}^{C \times 1 \times i \times i}$ denotes the $i \times i$ depth-wise convolution with the number of output channels C. $\alpha_i \in \mathbb{R}^{C \times 1 \times 1 \times 1}$ denotes the learnable vector. Let $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$, $\mathbf{Y}_i \in \mathbb{R}^{C \times H \times W}$, $\circ$ and $*$ denote the input feature map, output feature map, broadcast multiplication, and convolution operator, respectively. Each convolution path in Fig. 3 can be formulated as follows:

$$\begin{aligned} \mathbf{Y}_i &= \alpha_i \circ \mathbf{D}_i * \mathbf{X} \\ &= \mathbf{K}_i * \mathbf{X} \end{aligned} \tag{3}$$

where $\mathbf{K}_i$ denotes the weights of the new convolution kernel with kernel size i. Specially, $\mathbf{D}_i^{\mathrm{DG}}$ and $\mathbf{D}_i^{\mathrm{HV}}$ are generated by the combination of w and g, where DG and HV represent the two directions. After converting different convolution paths, we get $\mathbf{K}_3^{\mathrm{DG}}$, $\mathbf{K}_3^{\mathrm{HV}}$, and $\mathbf{K}_7$. The 3×3 depth-wise convolution is zero-padded and converted to 7×7 depth-wise convolution.

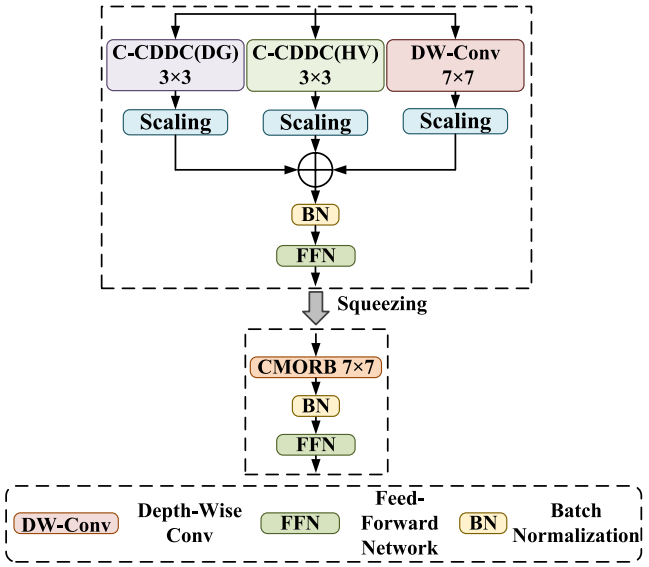


Fig. 3. Illustration of CMORB. Three branches are present within the block, a C-CDDC (DG) with a 3×3 convolution kernel, a C-CDDC (HV) with a 3×3 convolution kernel, and a 7×7 depth-wise convolution where 3 and 7 refer to the size of the convolution kernel. The three branches are equivalently converted into a normal single convolution branch 7×7 in the training and inference phases.

[1]More derivation details about (1) can be found in the supplementary material.

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 05,2024 at 01:19:02 UTC from IEEE Xplore. Restrictions apply.

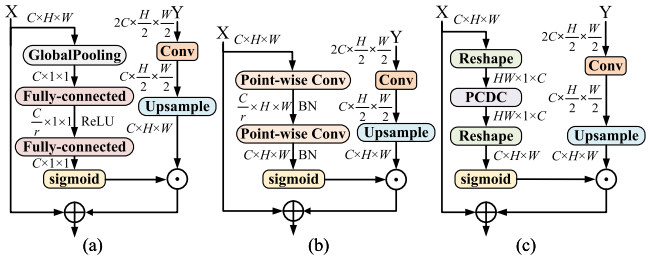


Fig. 4. Architecture of channel attention modulation modules for feature fusion. (a) GCAM [27], [30]. (b) PCAM with channel compression [26]. (c) Full spatial resolution FFMM with PCDC block. X and Y denote the low-level and high-level features, respectively. $\oplus$ and $\odot$ represent element-wise addition and element-wise multiplication, respectively.

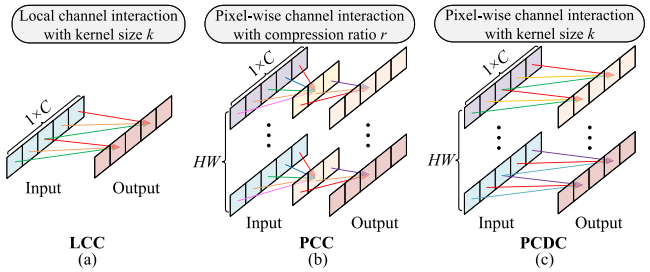


Fig. 5. Architecture of different channel attention mechanisms for enhancing target features. (a) LCC [35] performs channel attention calculation by employing a 1-D convolution with kernel size k on the channel dimension without channel compression. (b) PCC [27] conducts channel attention calculation by 1×1 normal convolution with compression ratio r on channel dimensions without spatial dimension compression. (c) PCDC calculates channel attention using depth-wise convolution with kernel size $k = 3$, applied to the full spatial resolution of the input without compressing in both spatial and channel dimensions. The feature map size is also presented after each step of the operation. i.e., C, H, W, and r denote the channel number, height, width, and compression ratio, respectively.

Then, the multiple convolution kernels are added together to generate the final 7×7 depth-wise convolution. The formulation can be denoted by

$$\begin{aligned}\mathbf{Y} &= \mathbf{Y}_7 + \mathbf{Y}_3^{\mathrm{DG}} + \mathbf{Y}_3^{\mathrm{HV}} \\ &= \mathbf{K}_7 * \mathbf{X} + z\left(\mathbf{K}_3^{\mathrm{DG}}\right) * \mathbf{X} + z\left(\mathbf{K}_3^{\mathrm{HV}}\right) * \mathbf{X} \\ &= \left(\mathbf{K}_7 + z\left(\mathbf{K}_3^{\mathrm{DG}}\right) + z\left(\mathbf{K}_3^{\mathrm{HV}}\right)\right) * \mathbf{X} \\ &= \mathbf{K}_7' * \mathbf{X} \end{aligned} \tag{4}$$

where z and $\mathbf{Y}$ denote zero-padded and fused feature maps of the multiple convolution paths, respectively. $\mathbf{K}_7'$ denotes the weight of the transformed single path depth-wise convolution kernel.

B. Pixel-Wise Cross-Channel Depth-Wise Convolution-Based Feature Fusion

Generally, the structural details are rich in the low layer which is beneficial for localization and small target tracking, while deep layer features are of high-level semantics which facilitate classification and large target tracking. The gradual aggregation of cross-layer feature maps in a top-down manner can effectively improve the performance of tracking multiscale targets. Existing works [24], [25] have successfully incorporated feature fusion into their tracking networks and achieved notable performance improvements. However, infrared UAV small targets are submerged by the background in the deep layer, making it difficult to effectively leverage the highly discriminative features by directly fusing the shallow layer and deep layer features.

To solve the above problems, we propose an FFMM by a novel full spatial resolution pixel-wise channel attention mechanism. The architecture of the FFMM is shown in Fig. 4(c). With the bottom-up modulation path, the spatial information of the low-level feature map can be accurately propagated to the deep layer which highlights the semantic features of small targets while suppressing irrelevant background information in the deep layer, leading to a more efficient and effective feature fusion process. Current channel attention mechanisms popularly model the feature importance using channel compression, which, inevitably leads to information loss [35], especially for small-scale targets. Therefore, we construct a novel FSRCA mechanism, to preserve crucial spatial properties while avoiding channel dimension compression. Within the FSRCA, we construct an effective pixel-wise cross-channel depth-wise convolution (PCDC) block with channel interaction to emphasize the details of IRUTs without squeeze-and-excitation (SE) operations while maintaining low-computational complexity.

Given below are more details about the proposed channel attention mechanism. Classical channel attention, such as SE attention [30], adopts one global pixel to represent the importance of each channel. Although it greatly reduces parameters, it introduces inaccurate channel importance for small targets due to the compression of spatial and channel information. The global channel attention modulation (GCAM) [27] based on SE attention is shown in Fig. 4(a). Dai et al. [26] proposed the improved pixel-wise channel attention modulation (PCAM), as shown in Fig. 4(b). Their methods adopt point-wise convolution to achieve channel interaction and then employ channel compression to reduce the model complexity. However, channel compression weakens small target features [35].

Small in scale, IRUTs are more sparse in distribution. Hence, we argue that the pixel-wise attention mechanism without spatial and channel compression is more appropriate for highlighting the IRUTs. As shown in Fig. 5(c), we construct the PCDC using pixel-wise cross-channel interaction in a full spatial resolution feature map, effectively avoiding the loss of small target information and obtaining the pixel-by-pixel channel importance. Compared with local cross-channel convolution (LCC) [35] shown in Fig. 5(a), our attention mechanism prevents simultaneous compression in both spatial and channel dimensions, thereby better preserving the features of small targets. As shown in Fig. 5(b), while pixel-wise cross-channel convolution (PCC) [27] avoids spatial compression, it still employs 1×1 convolution to compress channels, which inevitably results in the loss of small target information. Besides, in contrast to the PCC, where the same convolution kernel is applied to all spatial locations, our method employs different convolution kernels at each spatial location for the model to adaptively adjust the importance of channels with spatial perception.

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 05,2024 at 01:19:02 UTC from IEEE Xplore. Restrictions apply.

Assume the input feature map is $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$. We change its shape to $HW \times 1 \times C$. Then we employ depth-wise convolution with kernel size k to make local channel interaction. Finally, we reshape the dimension to the original. The contextual modulation weight $G(\mathbf{X}) \in \mathbb{R}^{C\times H\times W}$ can be calculated by

$$G(\mathbf{X}) = \sigma(\text{Reshape}(\text{DWConv}(\text{Reshape}(\mathbf{X})))) \tag{5}$$

where DWConv and σ mean the depth-wise convolution and sigmoid function, respectively. After high-level feature map upsampling, small target features are ambiguous since upsampling operations, such as bilinear interpolation, have no learnable parameters. Therefore, we multiply the deep-level feature maps by the weight map generated from FSRCA which contains more spatial details and facilitates the perception of the small target in the deep layer. Given the shallow layer feature $\mathbf{X}$ and the deep layer feature $\mathbf{Y}$, the fused feature map $\mathbf{O}$ can be obtained by

$$\mathbf{O} = \mathbf{X} + G(\mathbf{X}) \odot \text{Upsample}(\text{Conv}(\mathbf{Y})) \tag{6}$$

where $\odot$ and Conv denote element-wise multiplication and convolution, respectively.

We also show two existing mainstream attention modulation modules GCAM and PCAM in Fig. 4(a) and (b) for comparison with our method. Given the low-level feature $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, the SE attention can be formulated as

$$H(\mathbf{X}) = \sigma(\mathbf{W}_2\delta(\mathbf{W}_1(\text{AvgPool}(\mathbf{X})))) \tag{7}$$

and the channel attention in PCAM can be written as

$$F(\mathbf{X}) = \sigma(\text{BN}(\text{PWConv}_2(\delta(\text{BN}(\text{PWConv}_1(\mathbf{X})))))) \tag{8}$$

where σ, δ, AvgPool, BN, PWConv, and $\mathbf{W}$ denote the sigmoid function, rectified linear unit (ReLU), average pooling, batch normalization, point-wise convolution, and fully connected layer, respectively. $\mathbf{W}_1 \in \mathbb{R}^{(C/r)\times C}$, $\mathbf{W}_2 \in \mathbb{R}^{C\times(C/r)}$, $\text{PWConv}_1 \in \mathbb{R}^{(C/r)\times C\times 1\times 1}$, and $\text{PWConv}_2 \in \mathbb{R}^{C\times(C/r)\times 1\times 1}$, where r means the channel compression ratio. Give the deep-level feature map $\mathbf{Y} \in \mathbb{R}^{2C\times(H/2)\times(W/2)}$, then the output feature maps $\mathbf{L}$ and $\mathbf{R}$ of GCAM and PCAM can be respectively computed by

$$\mathbf{L} = \mathbf{X} + H(\mathbf{X}) \odot \text{Upsample}(\text{Conv}(\mathbf{Y})) \tag{9}$$

$$\mathbf{R} = \mathbf{X} + F(\mathbf{X}) \odot \text{Upsample}(\text{Conv}(\mathbf{Y})). \tag{10}$$

C. Cross-Attention-Based Updatable Feature Interaction Module

Accurately enhancing IRUT features in the search frame is beneficial for detecting UAV targets. SiamFC [4], SiamRPN [6], and SiamRPN++ [7] utilize local cross correlation for the interaction between the template and search frame. However, leveraging these feature interaction methods will blur the target boundaries and can only roughly perceive the search frame target location [36]. Besides, the appearance, size, illumination, and background changes of the UAV during the movement make the tracking task more challenging.

For the IRUT tracking task, we devise a CUFIM which utilizes pixel-by-pixel similarity calculation to fully leverage the information of small targets and enable the model to adapt to the changes of the target state by introducing a multitemplate. Compared with local correlation operation, Transformer-based feature interaction can better resist similar target interference [37]. With pixel-by-pixel global relationship modeling, the small target information of the template can be better utilized to improve the awareness of the search branch for the target. Previous works, such as [1] and [2], have introduced a global search strategy into the tracking framework which shows superior performance in the long-term tracking task. However, these methods are susceptible to drifting toward interference as they only utilize the initial template and are insufficient to extract discriminative IRUT features. Thanks to long-distance dependency modeling, our tracker can effectively resist interference from similar targets under the global search paradigm. Furthermore, the model's ability to identify interferences is further enhanced by excavating temporal association with multitemplate.

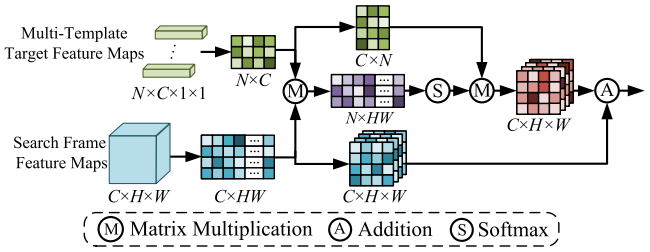


Fig. 6. Architecture of the proposed CUFIM. The initial template and extra high-quality multitemplate are utilized as input along with the search frame to the feature interaction module. This module is intended to improve the tracking performance of the model under conditions of changes in the state of the target and the background by integrating multitemplate information into the search frame.

The UAV target is dynamically deformed during the movement and the surrounding background may change greatly compared to the initial frame. The initial template can hardly offer sufficient information in this situation. To cope with this problem, we introduce extra high-quality templates in the following tracking process to help the model better locate the target in the search frame. The additional templates can perceive the change in the state of the target and thus can provide more accurate target-related information to the search frame. By utilizing these templates, our tracker is able to precisely estimate the location and scale of UAV targets.

The constructed feature interaction module is shown in Fig. 6. $F_{\text{tmp}} \in \mathbb{R}^{N\times C\times 1\times 1}$ denotes the target region-of-interest (ROI) features of the multitemplate and $F_{\text{src}} \in \mathbb{R}^{C\times H\times W}$ denotes the search frame features. N, C, H, and W represent the number of templates, the channel dimension, height, and width, respectively. We first transform the multitemplate target feature maps as the key $K \in \mathbb{R}^{N\times C\times 1\times 1}$ and value $V \in \mathbb{R}^{N\times C\times 1\times 1}$. Then we convert the search feature into query $Q \in \mathbb{R}^{C\times H\times W}$. The calculation of the similarity matrix is performed by the matrix multiplication of Q and K after reshaping query and key dimension to $Q \in \mathbb{R}^{C\times HW}$ and $K \in \mathbb{R}^{N\times C}$. After the similarity calculation, we acquire the spatial pixel-by-pixel similarity matrix. Then, we calculate the cross-attention feature of the search frame by multiplying the similarity matrix

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 05,2024 at 01:19:02 UTC from IEEE Xplore. Restrictions apply.

with the V after reshaping the dimension to $V \in \mathbb{R}^{C\times N}$. Finally, the obtained feature map is added to F_{src} which enhances the target region of the search frame. The equation for feature fusion can be written as

$$F'_{\text{src}} = V \otimes \text{Softmax}\left(\frac{K \otimes Q}{\sqrt{d}}\right) + F_{\text{src}} \tag{11}$$

where $\otimes$ denotes matrix multiplication, F'_{src} is the search frame feature map with an enhanced target region, and d represents the dimension of the key.

D. Multiscale Search Branches of IRUTs

Current methods [21], [38] only utilize feature maps from the last stage of the network which are not conducive to the detection of multiscale IRUTs. Small targets only occupy a small part of the image and the details of these targets are lost with several downsampling operations. In this work, we construct three global tracking branches based on the FFMM and CUFIM to overcome the limitation by only applying a single tracking branch.

Our goal is to effectively utilize different hierarchical-level features and precisely propagate deep semantic information to shallow layers for better tracking performance of multiscale targets. We construct a top-down path that sufficiently exploits the features of different layers and meanwhile introduces three tracking branches for detecting multiscale targets. Specifically, we leverage the output feature maps from the last block of each stage. As shown in Fig. 1, we denote the output feature maps from stages 1 to 4 as F_1, F_2, F_3, and F_4, respectively and the fused feature maps from FFMM as G_1, G_2, and G_3, respectively. We first upsample F_4 and get the coarse spatial detail feature map. With our FFMM, the feature map can be refined by a bottom-up modulation path which improves the spatial details perception in the deep layer features. Then the refined feature map is merged with F_3 by element-wise addition and get G_1. This iterative process will continue until the last G_3 is generated. As a result, the feature maps G_1, G_2, and G_3 all incorporate rich semantic information from deep layer features which is beneficial for subsequent tracking of multiscale IRUTs.

IRUTs are usually small in scale and are frequently interfered by complex backgrounds, so they can easily be occluded by backgrounds. Besides, the location where it appears in subsequent frames is not necessarily near the location where it appeared the last time. Thus local search-based methods can hardly continuously track the target. Instead, we constructed three global tracking branches which give the model the ability to recapture the target and resist the influence of the complex background making the tracking framework more suitable for infrared UAV tracking tasks.

E. Anchor-Free Detection Head

In multiscale datasets, the design of the anchor-based head leads to the imbalance problem of sampling large targets and small targets which makes the model pay less attention to small targets. An anchor-free head can alleviate the problem of sample imbalance. Besides, with the anchor-free head, we do not need to set a number of hyperparameters such as the IoU threshold, the number of anchor boxes, and the aspect ratio of boxes. In this work, we choose to adjust the anchor-based YOLOv5 [39] detection head to generate an anchor-free detection head.

To be specific, the head consists of a classification branch, an objectness branch, and a regression branch. The first two branches find out where the target is most likely to be and the third branch regresses the size of the predicted box. The head is shared by the three tracking branches. For selecting positive samples, we first map the center point pixel of the target backing to the feature map of the three tracking branches. The mapped grid and the two grids closest to it are considered positive samples. If the center of the mapped ground truth falls on the center of the current grid, only that grid is considered a positive sample. The regression task is to estimate the target box size and the distance from the mapped point on the feature map to the center of the target. Specifically, the bounding box is determined by (b_x, b_y, b_w, b_h) denoting the target center and target size. Given the coordinates position (c_x, c_y) in the feature map and the output (t_x, t_y, t_w, t_h) from the predicted head, the calculation can be defined as

$$\begin{aligned} b_x &= (2 \times \sigma(t_x) - 0.5) + c_x \\ b_y &= \left(2 \times \sigma\left(t_y\right) - 0.5\right) + c_y \\ b_w &= e^{t_w} \\ b_h &= e^{t_h} \end{aligned} \tag{12}$$

where σ denotes the sigmoid function.

Since the features from different layers are responsible for tracking targets of different scales, we set different size ranges for three tracking branches. If the target box area is between n_{i-1} and n_i, it is considered as the positive sample on feature level i. In this work, n_0, n_1, n_2, and n_3 are set as 0, 196, 1024, and 4096. In the inference phase, the final confidence score is multiplied by the predicted classification score and objectness score. In this way, it can alleviate the problem of classification and localization misalignment to some extent. The loss function is defined as

$$L = \frac{1}{N_{\text{pos}}}\left\{\sum_j L_{\text{cls}}\left(s_j, \left[gt_j^l \geq 1\right]\right) + \sum_j \left[gt_j^l \geq 1\right] L_{\text{reg}}\left(t_j, gt_j^b\right) + \sum_j L_{\text{obj}}\left(o_j, \left[gt_j^l \geq 1\right]\right)\right\} \tag{13}$$

where N_{pos} indicates the number of positive samples, L_{cls} and L_{obj} are the focal loss [40]. L_{reg} is the GIoU loss [41]. j defines the index of the selected samples, gt_j^l and gt_j^b are the class label and location target, s_j is the predicted classification score, t_j is the predicted offset, o_j is the predicted objectness score, and $[gt_j^* \geq 1]$ is the indicator function, being 1 if $gt_j^* \geq 1$ and 0 otherwise.

IV. Experiments

In this section, we conduct experiments to validate the effectiveness of our method in comparison with other state-of-the-art (SOTA) baseline methods. We first introduce our

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 05,2024 at 01:19:02 UTC from IEEE Xplore. Restrictions apply.

datasets and the experiment setups we employ to train and test all methods, and then present the quantitative and qualitative experiment results. Finally, we design some ablation studies to verify the effectiveness of the proposed components.

A. Datasets

We select ten infrared image sequences, each of which represents common challenges encountered in anti-UAV scenarios, such as heavy clouds and similar backgrounds. Seqs. 1–5 and 7 are collected from public datasets and can be accessed online [42], [43]. The rest are collected by ourselves and carefully labeled. We use a total of 19 392 images that cover targets of various scales. The division ratio of the training and testing sets is set to 7:3. More details about the datasets can be found in the supplementary material.

B. Evaluation Metrics and Experimental Setups

In the training phase, we employ the SGD optimizer with a momentum of 0.9 and a weight decay of 0.0001. We use a cosine learning rate schedule, where warmup epochs are 50 and the base learning rate is set to 0.003. The total training epoch and batch are 250 and 16, respectively. Each training pair consists of a template and search frame from the same sequence with a random interval smaller than 200. We use random flipping for data augment. The network is trained from scratch without loading pretraining parameters. We implemented it on a server with an NVIDIA GeForce RTX 3090 GPU accelerated by CUDA 11.1. For a software implementation, we use Python 3.8 and Pytorch 1.8.1. In the inference phase, we take the target region from the first frame as the initial template. Additionally, the dynamic template number N is set to 2. We send the full image to the network and resize the image size to 384 $\times$ 384 without cropping. We update the dynamic templates only when the confidence score of the predicted target is higher than 0.9. The high threshold prevents the model from being polluted by inaccurate templates. The oldest template will be removed when the maximum number of dynamic templates is exceeded. We set the confidence score threshold to 0.4. In cases where the confidence score τ falls below this threshold, the model will predict the absence of any targets in the current frame.

We select the CNN-based methods SiamRPN-LT [6], GlobalTrack [1], Ocean [38], Unicorn [2], and SiamCAR [21] as well as the transformer-based methods GRM [12], AiATrack [11], and OSTrack [10] for comparison with our method. We also choose the IRUT tracking method SiamYOLO [20] for comparison. Among them, SiamCAR and Ocean belong to anchor-free tracking methods; Unicorn, GlobalTrack, and SiamYOLO are global search-based tracking methods. Other methods are local search-based methods. AiATrack and Ocean are tracking methods that introduce an online updating mechanism. Compared with the original version, SiamRPN-LT has a larger search range to meet the needs of long-term tracking.

In detail, for SiamRPN-LT, its backbone is AlexNet; for OSTrack, its backbone is ViT-Base; for GRM, its backbone is ViT-Large; for Ocean, SiamCAR, AiATrack, and GlobalTrack, their backbones are ResNet-50; for Unicorn, its backbone is ConvNeXt-Tiny; and for SiamYOLO, its backbone is Darknet-53. All experiments are trained on the same infrared UAV datasets and tuned to the optimal.

We use the precision plot [44], success plot [44], normalized precision (NPrecision) plot [45], and state accuracy (SA) [18] to evaluate the performance of each method. The SA is commonly used in long-term tracking evaluation, defined as

$$\mathrm{SA} = \sum_t \frac{\mathrm{IoU}_t \times [v_t > 0] + m_t \times (1 - [v_t > 0])}{T} \tag{14}$$

where T denotes the frames number of each sequence. The tracker's predicted m is used to measure the SA. v_t is the ground-truth existence flag of the IRUT. $[v_t > 0]$ is the indicator function, being 1 if $v_t > 0$ and 0 otherwise. The mSA calculates the average SA of all sequences. The precision plot calculates the percentage that the center distances between predicted and ground-truth bounding boxes are lower than a given threshold. We use the precision at a threshold of 5 pixels as a representative score per tracker which is suitable for evaluating the tracking performance of small IRUTs [46]. The NPrecision is calculated by normalizing the precision over the size of the ground-truth bounding box. We use the area under the curve (AUC) as a representative score for ranking each tracker in the NPrecision plot. The success plot calculates the percentage that the IoU between predicted and ground-truth bounding boxes is greater than a given threshold. We use the AUC as a representative score for ranking each tracker.

We also compare our method with SOTA methods on the public infrared Anti-UAV target tracking dataset [18] and infrared single object tracking dataset LSOTB-TIR [47]. These experiment results are presented in the complementary material.

C. Quantitative Results

1) Overall Performance: Table I shows the quantitative results of the nine aforementioned comparison methods and ours in ten infrared image sequences. From a comprehensive perspective, it can be seen that our proposed method performs better on all sequences. Especially for Seqs. 2, 9, and 10, our method exhibits a significant advantage over other methods which effectively demonstrates its effectiveness in tracking small and weak targets within complex backgrounds. We average the SA across all sequences to obtain the mSA and our method surpasses Unicorn by 6.6% in mSA.

Furthermore, we present the success plot, precision plot, and NPrecision plot in Fig. 7. A larger area under the curve means better tracking performance. It can be observed that our method is consistently superior in location accuracy and estimation of the scale compared with other methods. In particular, our method outperforms the second-best method Unicorn by 6.6% in success, by 1.6% in precision, and by 5.5% in NPrecision.

We also show the inference speed of each method with frames per second (FPS) in Table I. The FPS we analyze below is averaged over all sequences. The transformer-based method

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 05,2024 at 01:19:02 UTC from IEEE Xplore. Restrictions apply.

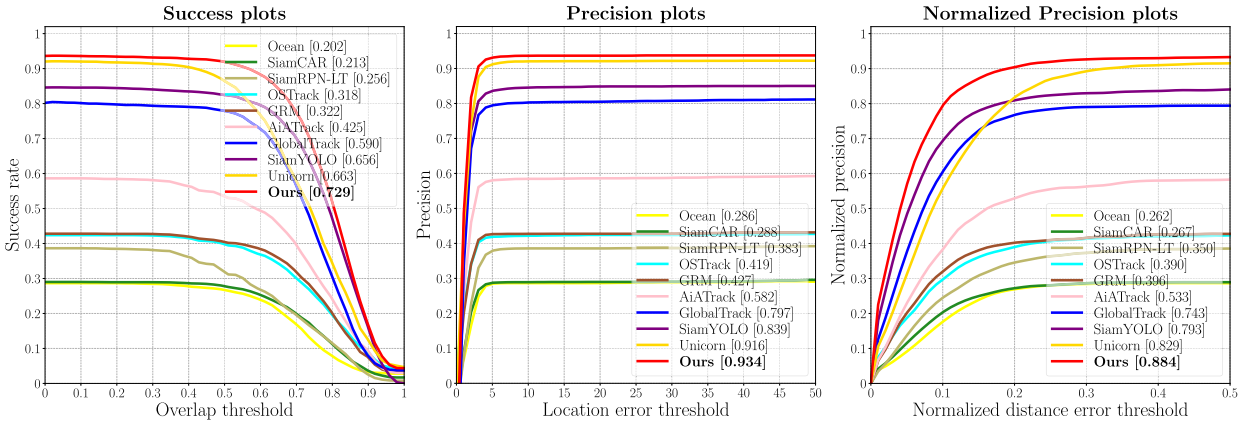


Fig. 7. Success plot, precision plot, and NPrecision plot of ten trackers.

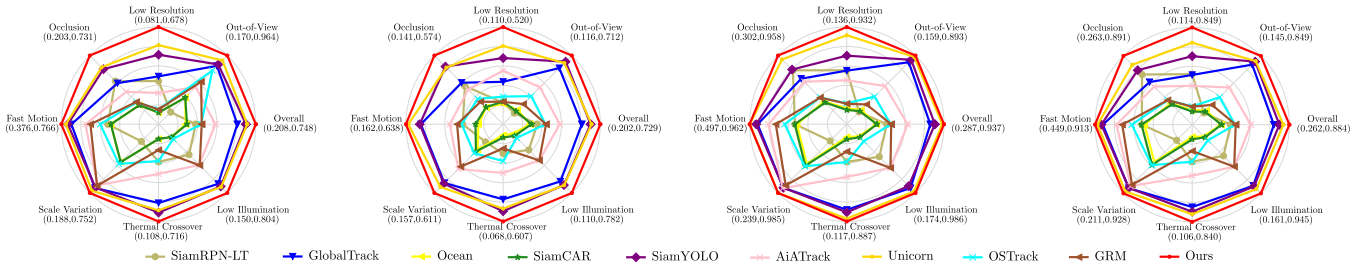


Fig. 8. Performance comparison of different trackers using attribute-based score radar plots. From left to right, they are the mSA plot, success plot, precision plot, and NPrecision plot, respectively.

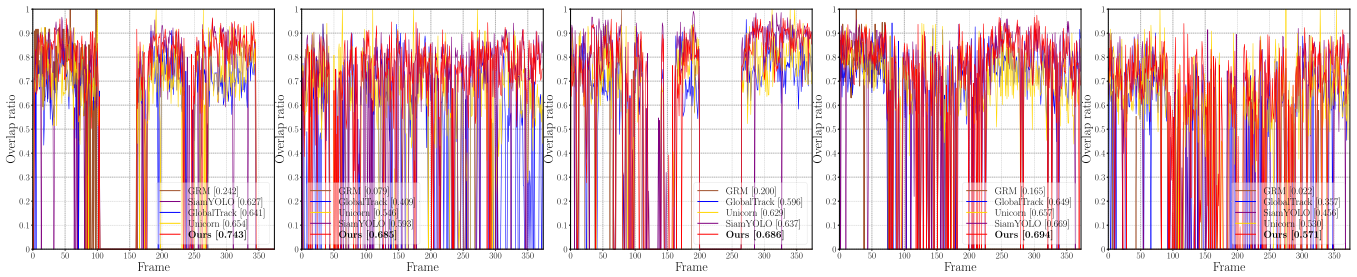


Fig. 9. Overlap ratio plots of five methods on five challenge sequences from Seqs. 1 to 5. The performance is represented by the average IoU. In Seq. 1, the target is out of view from the 105th to 161st frames, as well as from the 347th to 374th frames. In Seq. 2, the target is out of view from the 45th to 49th frames. In Seq. 3, the target is out of view at the 69th, 70th, 91st, 92nd, 114th, and 115th frames, and from the 125th to 131st frames, as well as from the 201st to 265th frames. In Seq. 4, the target is out of view at the 121st and 146th frames.

GRM achieves 14.0 frames/s, illustrating the complexity of its network design. For GlobalTrack and Unicorn, they reach 9.6 and 28.9 frames/s, respectively. In contrast, our method can achieve real-time tracking of 34.5 frames/s attaining a good balance between tracking performance and inference speed.

2) Performance Comparison of Attribute-Based Trackers: To conduct a more comprehensive comparison of different trackers, we evaluate their performance in various attributes to comprehensively analyze the overall tracking capability of each tracker. These attributes include out-of-view (OV), OC, FM, scale variation (SV), low illumination (LI), thermal crossover (TC), and low resolution (LR). To better evaluate the performance of different trackers for small targets, we modify the scale definition. Specifically, we redefined the small-scale size from the original threshold of 400 to 80 which follows the definition in [48]. Anti-UAV [42] datasets include attribute annotations for each video, and for our datasets, we meticulously labeled each frame of the sequence based on the description of each attribute.

Fig. 8 illustrates the mSA, success, precision, and NPrecision of all trackers under seven challenge attributes. Compared with other trackers, our tracker shows superior performance in different challenge scenarios. This is mainly attributed to the proposed CMORB and FFMM, which effectively enhance the representation ability of IRUTs with enhanced contextual information and multilevel hierarchical feature representation. Compared to the global search-based methods GlobalTrack, SiamYOLO, and Unicorn, our tracker can better locate the target and estimate the size of the predicted bounding box. The main reason is that the proposed online updating module integrates rich historical information improving the model's perception of the changes in the state of the target.

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 05,2024 at 01:19:02 UTC from IEEE Xplore. Restrictions apply.

TABLE I

QUANTITATIVE COMPARISONS OF THE PROPOSED METHOD AND OTHER METHODS. THE BEST THREE RESULTS ARE SHOWN IN RED, GREEN, AND BLUE FONTS, RESPECTIVELY

# Seq.	Metrics	SiamRPN-LT	GlobalTrack	Ocean	SiamCAR	SiamYOLO	AiATrack	Unicorn	OSTrack	GRM	Ours
1	SA	0.011	0.661	0.301	0.249	**0.700**	0.484	**0.716**	0.594	0.347	**0.799**
	Success	0.011	0.644	0.293	0.242	**0.682**	0.473	**0.698**	0.580	0.340	**0.780**
	Precision	0.011	0.864	0.371	0.317	**0.872**	0.592	**0.909**	0.704	0.400	**0.933**
	NPrecision	0.015	0.787	0.338	0.288	**0.801**	0.546	**0.817**	0.653	0.376	**0.877**
	FPS	31.6	9.2	**44.0**	**55.2**	**60.3**	35.8	28.9	43.4	15.2	35.1
2	SA	0.100	0.415	0.009	0.009	**0.596**	0.491	**0.544**	0.101	0.083	**0.689**
	Success	0.094	0.402	0.009	0.008	**0.580**	0.479	**0.528**	0.098	0.081	**0.671**
	Precision	0.131	0.608	0.008	0.008	**0.781**	0.659	**0.813**	0.131	0.112	**0.912**
	NPrecision	0.119	0.532	0.010	0.010	**0.709**	0.584	**0.691**	0.119	0.103	**0.827**
	FPS	13.5	9.7	44.1	**55.0**	**60.0**	39.2	29.1	**46.9**	14.9	35.3
3	SA	0.209	0.668	0.144	0.126	**0.691**	0.306	**0.688**	0.324	0.260	**0.738**
	Success	0.205	0.652	0.141	0.123	**0.679**	0.301	**0.673**	0.318	0.251	**0.720**
	Precision	0.288	**0.845**	0.152	0.133	0.825	0.363	**0.848**	0.381	0.304	**0.861**
	NPrecision	0.252	0.786	0.147	0.129	0.769	0.336	**0.773**	0.355	0.283	**0.815**
	FPS	20.1	9.6	43.8	**55.6**	**61.4**	39.5	28.9	**47.0**	15.7	35.1
4	SA	0.164	0.646	0.158	0.169	**0.665**	0.283	**0.653**	0.170	0.170	**0.693**
	Success	0.152	0.629	0.154	0.165	**0.652**	0.276	**0.636**	0.165	0.165	**0.676**
	Precision	0.206	**0.871**	0.212	0.217	0.822	0.370	**0.906**	0.220	0.212	**0.917**
	NPrecision	0.185	0.799	0.189	0.198	0.772	0.333	**0.796**	0.198	0.193	**0.808**
	FPS	11.4	9.7	44.0	**55.6**	**59.4**	39.2	28.9	**46.6**	15.7	35.0
5	SA	0.023	0.357	0.040	0.020	**0.456**	0.043	**0.530**	0.057	0.022	**0.571**
	Success	0.022	0.347	0.023	0.019	**0.443**	0.0417	**0.514**	0.055	0.021	**0.555**
	Precision	0.040	0.525	0.043	0.029	**0.637**	0.064	**0.811**	0.083	0.032	**0.819**
	NPrecision	0.032	0.455	0.034	0.026	**0.563**	0.056	**0.680**	0.070	0.027	**0.713**
	FPS	10.5	9.8	44.1	**55.9**	**59.6**	39.4	29.1	**46.7**	13.0	35.1
6	SA	0.378	**0.733**	0.699	0.723	0.731	0.725	**0.733**	0.723	**0.736**	**0.773**
	Success	0.320	**0.713**	0.679	0.702	0.711	0.704	**0.713**	0.703	**0.716**	**0.754**
	Precision	0.464	0.987	**0.998**	**1.000**	0.978	**1.000**	**1.000**	**0.999**	0.995	**1.000**
	NPrecision	0.424	**0.909**	0.888	0.904	0.904	0.892	**0.906**	0.894	0.902	**0.934**
	FPS	14.9	9.6	**44.2**	**56.2**	**63.0**	38.9	29.0	**44.2**	13.0	35.3
7	SA	0.687	0.856	0.152	0.237	**0.866**	0.600	**0.872**	0.582	0.448	**0.877**
	Success	0.627	0.836	0.146	0.227	**0.847**	0.584	**0.850**	0.567	0.435	**0.857**
	Precision	0.734	**0.992**	0.178	0.266	0.942	0.664	**0.994**	0.630	0.516	**1.000**
	NPrecision	0.712	**0.947**	0.167	0.256	**0.914**	0.654	**0.947**	0.631	0.491	**0.949**
	FPS	21.1	9.6	43.9	**55.8**	**60.6**	38.5	28.9	**46.7**	13.0	33.0
8	SA	0.324	**0.816**	0.177	0.218	0.806	0.708	**0.815**	0.229	0.744	**0.817**
	Success	0.285	**0.796**	0.172	0.213	0.787	0.691	**0.795**	0.224	0.726	**0.798**
	Precision	0.366	**0.998**	0.215	0.255	**0.962**	0.862	**1.000**	0.272	0.908	**1.000**
	NPrecision	0.345	**0.943**	0.204	0.244	0.911	0.803	**0.931**	0.257	0.850	**0.935**
	FPS	14.3	9.5	43.7	**55.8**	**62.0**	38.5	28.8	**46.8**	13.0	31.3
9	SA	0.402	0.523	0.050	0.050	**0.613**	0.137	**0.612**	0.090	0.100	**0.786**
	Success	0.300	0.507	0.047	0.048	**0.597**	0.131	**0.594**	0.086	0.095	**0.766**
	Precision	0.614	**0.772**	0.114	0.083	0.781	0.286	**0.917**	0.200	0.219	**1.000**
	NPrecision	0.532	0.650	0.092	0.067	**0.720**	0.223	**0.773**	0.147	0.156	**0.920**
	FPS	16.4	9.5	43.9	**55.7**	**63.0**	39.4	29.0	**48.7**	13.0	35.0
10	SA	0.565	0.383	0.366	0.392	**0.600**	0.591	**0.653**	0.397	0.404	**0.734**
	Success	0.545	0.373	0.354	0.381	**0.583**	0.573	**0.633**	0.385	0.392	**0.714**
	Precision	**0.976**	0.512	0.574	0.574	0.793	0.959	**0.962**	0.574	0.568	**0.997**
	NPrecision	**0.813**	0.469	0.495	0.495	**0.712**	0.797	**0.813**	0.499	0.503	**0.899**
	FPS	**52.5**	9.5	44.2	**56.2**	**61.1**	39.2	28.8	48.8	13.0	35.1

Due to the long imaging distance for UAVs, the targets are always small in scale and can be easily occluded. Under such conditions, many trackers struggle to achieve satisfactory performance for extracting features of weak and small targets. Our tracker shows outstanding performance in the OC and LR scenarios in terms of mSA and success. Specifically, in LR scenarios, our method demonstrates a 12.6% improvement in mSA, 10.1% in success, 7.9% in precision, and 13.3% in NPrecision and obtains 12.5%, 10.0%, 5.5%, and 11.5% increments in terms of the mSA, success, precision, and NPrecision in OC scenarios compared with the second-best method Unicorn.

The values of IRUTs sometimes are similar to that of the surrounding background which makes it difficult for the tracker to differentiate the target from the background accurately. Meanwhile, the target with LI can easily cause the target to blend into the background. In LI scenarios, our method showed 7.1% improvement in mSA, 8.8% in success, 2.2% in precision, and 6.1% in NPrecision over Unicorn and attained performance improvement in mSA of 8.2%, success of 8.0%, precision of 2.6%, and NPrecision of 6.6% in TC scenarios. These results show that our method can sufficiently extract the discriminative feature of IRUTs.

Our model is capable of effectively addressing the challenges posed by target deformation and complex background changes, thanks to the proposed CUFIM which introduces multiple templates effectively improving the target representation ability in the search frame branch. Compared with

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 05,2024 at 01:19:02 UTC from IEEE Xplore. Restrictions apply.

AiATrack which integrates the online updating mechanism, our method improves the performance by 8.5%, 21.3%, 6.8%, and 9.8% in terms of mSA, success, precision, and NPrecision in scenarios with scale variation.

The frequent disappearance and reappearance of UAV targets pose significant challenges to the tracking task. Overall, local search strategies tend to be less effective in comparison to global search strategies. Compared with three global search-based trackers, our tracker gains competitive performance in mSA, success, precision, and NPrecision. This shows that our tracker can effectively extract difference information between the target and the background while excavating semantic features from the deep layer and precisely transferring to the shallow layer to better identify the target and judge the absence of the target. Concretely, our method outperformed Unicorn, with a 6.9% improvement in mSA, 7.8% improvement in success, 1.8% improvement in precision, and 6.4% improvement in NPrecision.

FM is a common challenge in anti-UAV tracking tasks, where sudden UAV acceleration can cause a significant deviation in the target position between two adjacent frames. Compared with Unicorn, our method has a 3.5% improvement in mSA, 1.6% improvement in success, 0.3% improvement in precision, and 2.4% improvement in NPrecision. In such scenarios, employing a global search mechanism for IRUT tracking can be beneficial to better accuracy and robustness.

D. Qualitative Results

Fig. 9 shows overlapping curves on five challenging scenarios from Seq. 1 to 5. Compared with other methods, our method achieves the best performance in terms of the average IoU, validating the effectiveness of our tracker for the IRUT tracking task. Fig. 10 shows the infrared target tracking results of different methods from Seq. 1 to 5.

1) Out of View: In Seq. 1, the UAV target is out of view at the 134th frame, and both Unicorn and GlobalTrack mistakenly predict the presence of the target due to the weak feature extraction ability of these models which makes them easily interfered by distractors. In contrast, our tracker can precisely extract the discriminative features of the small and weak UAV targets. This is attributed to our proposed FFMM, which accurately leverages the semantic features of small targets in the deep layer while efficiently improving the representation ability of small targets in the shallow layer. When the target reappears in the 162nd frame, our tracker quickly recaptures it and exhibits superior location accuracy compared to other methods.

2) Small Scale: Seq. 2 shows the small and dim UAV target moves in the forest. In Seq. 2(c)–(e), the tracked UAV target only accounts for a small portion of the whole image, which makes it susceptible to blending into the forest background and increases the difficulty of the tracking task. GlobalTrack, SiamYOLO, and Unicorn can easily mis-track or only track the part of the target. In contrast, the proposed CMORB is capable of effectively capturing the difference between the target and the background along with rich contextual information for infrared small UAV targets, enabling our tracker to consistently track the small targets and resist similar target interference.

TABLE II
EFFECTIVENESS OF INTEGRATING CMORB, FFMM, AND CUFIM

Row	CMORB	FFMM	CUFIM	Metrics			
				mSA	Success	Precision	NPrecision
1	-	-	-	0.684	0.667	0.856	0.816
2	✓	-	-	0.710	0.692	0.887	0.841
3	-	✓	-	0.724	0.706	0.905	0.858
4	-	-	✓	0.705	0.687	0.885	0.837
5	✓	✓	-	0.735	0.717	0.917	0.869
6	✓	-	✓	0.722	0.704	0.905	0.856
7	-	✓	✓	0.736	0.717	0.922	0.872
8	✓	✓	✓	**0.748**	**0.729**	**0.934**	**0.884**

3) Occlusion: In Seq. 3, the UAV target is seriously occluded by the signal tower. In Seq. 3(a) and (b), the target is partially occluded, and both GlobalTrack and Unicorn can only provide rough estimations of the target's location and scale. Seq. 3(c)–(e) presents a more challenging situation where the target is completely occluded by the tower, and only our method is able to successfully track the target. Owing to CUFIM, our method can adapt to the changes in the target and its surrounding background.

4) Camera Move: In Seq. 4(c)–(e), the camera rotates the view to recapture the target when it is about to go out of sight. In Seq. 4(d), the camera move causes motion blur, resulting in significant changes in the shape of the target. Compared with other methods, our method can still better locate the target. In Seq. 4(a), GlobalTrack is interfered from the distractor, indicating that the tracker has difficulty acquiring accurate target features in complex backgrounds. Equipped with CMORB and FFMM, our tracker enjoys enhanced representation of the target feature, thus its ability to resist the influence of interfering targets.

5) Fast Motion: In Seq. 5(a) and (b), the UAV target moves rapidly with a significant displacement within 2 s. In this case, the local tracker GRM can hardly track the target when tracking failure occurs. Seq. 5(c)–(e) shows the scenario of the target moves in building backgrounds within one second. The target's brightness is almost identical to that of the background, presenting a difficulty for comparison methods to accurately track the target. In particular, the target is completely merged into the background in Seq. 5(d) and (e), and only our method is capable of precisely identifying its location. The insufficient contrast between the target and background severely undermines the performance of these trackers. In contrast, our method can effectively excavate fine-grained discriminative features of the IRUT, enabling the model to stably track the target.

E. Ablation Study

We perform ablation studies in terms of CMORB, FFMM, and CUFIM to verify the effectiveness of each of them. The results are shown in Table II. We adopt ConvNeXt-Tiny, FPN, and anchor-free detection heads to construct our baseline.

1) Effectiveness of Integrating All Sub-Component: Table II shows the quantitative results of our tracking method with/without the proposed CMORB, FFMM, and CUFIM. The results of rows 1 and 8 in Table II illustrate that compared with

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 05,2024 at 01:19:02 UTC from IEEE Xplore. Restrictions apply.

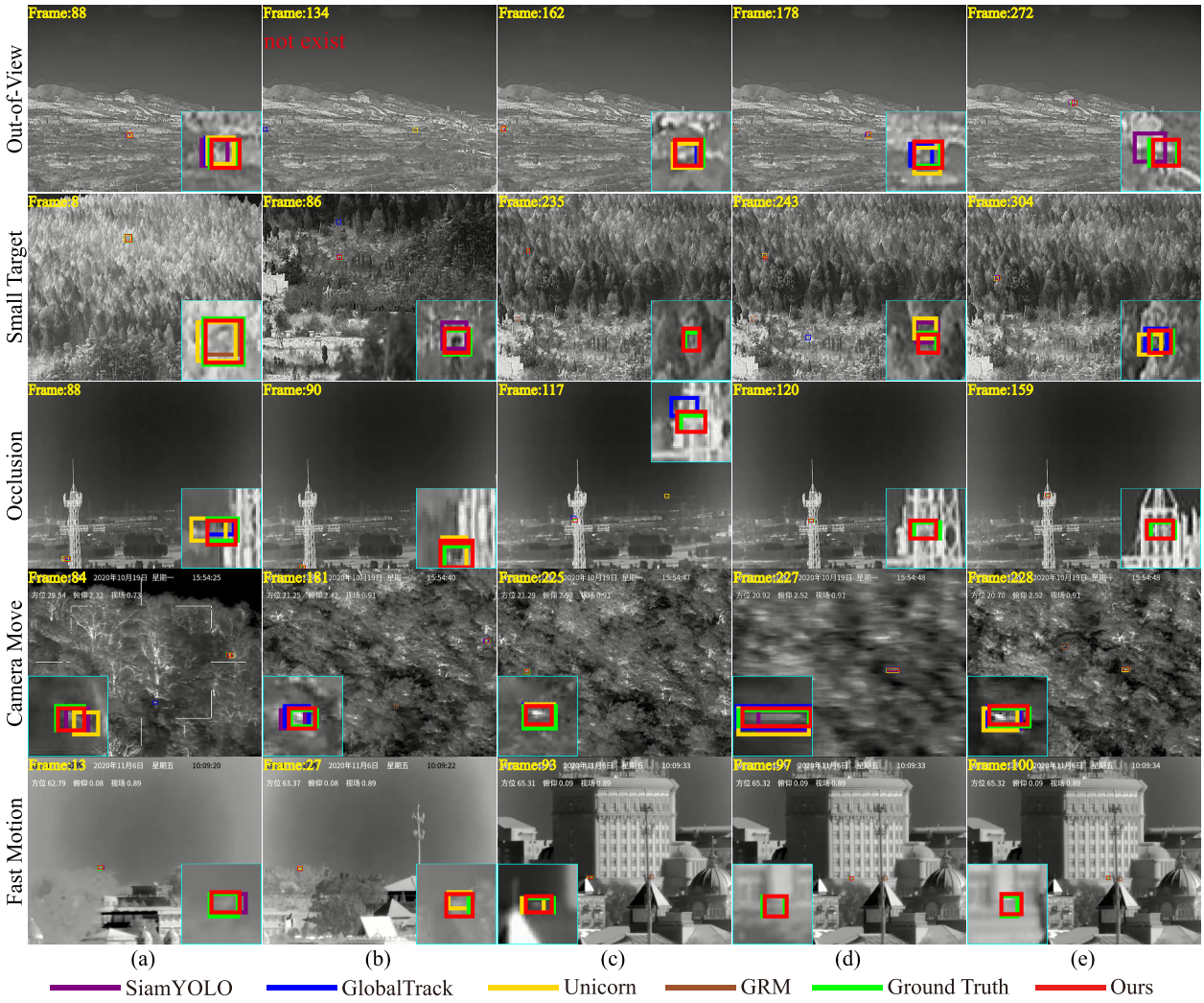


Fig. 10. Qualitative tracking results of five methods in terms of five attributes on five challenge sequences from Seqs. 1 to 5. Each row presents a sequence with five representative images. Close-ups are given for better visualization. (a)–(e) Five representative images in each sequence.

baseline, mSA, success, precision, and NPrecision increased by 6.4%, 6.2%, 7.9%, and 6.8%, respectively. We can clearly see that incorporating any of the proposed components can enhance the tracking performance compared to the baseline, suggesting that each of these techniques can improve the model's ability to track IRUTs under complex backgrounds. Integrating all components achieves better results than a single or paired combination, indicating that each component has complementary advantages.

2) Effectiveness of Integrating Online Re-Parameterization Building Block: As shown in rows 1 and 2 of Table II, changing each block of the backbone to CMORB brings an improvement of mSA by 2.6%, success by 2.5%, precision by 3.1%, and NPrecision by 2.5% which means the introduced contrast-based branches can effectively enhance backbone feature extraction capabilities for IRUTs. Compared with the single branch structure which is limited by the contribution of a single hierarchical feature map to the detection of targets, our proposed CMORB not only utilizes contrast difference information but also integrates convolution kernels of various sizes to achieve multireceptive field feature expressions, resulting in improved performance in extracting IRUT features.

Fig. 11 exhibits the feature maps of the convolution layer which is generated from the last block of stage 1 in the backbone network before and after adding contrast-based branches. It is evident that before adding these branches, the extracted features of the model were scattered and lacked focus on the target area. However, after employing two C-CDDC paths, the target features are significantly enhanced, with heightened contour information and a more focused target region. With these obvious target features, our method can more effectively transfer target information to the deep layers of the network. The visualization results well demonstrate that the proposed block is able to efficiently utilize the difference information between the target and the background.

Table III comprehensively shows the performance improvement brought by different branches. Compared with only introducing a 3×3 convolution branch, adding any contrast-based branch can bring better performance enhancement, indicating that the combination of both local feature and contrast intensity feature can better capture IRUTs in

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 05,2024 at 01:19:02 UTC from IEEE Xplore. Restrictions apply.

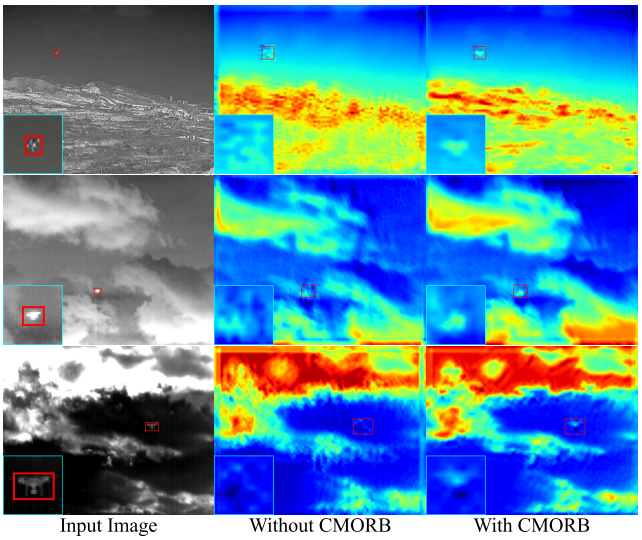


Fig. 11. Visualization of response maps from three representative sequences before and after applying CMORB. The red box indicates the tracking target. The first column shows the input images. The second and third columns display the response maps obtained without and with the proposed CMORB. The features of the target region are significantly enhanced after leveraging CMORB.

TABLE III

EFFECTIVENESS OF INTEGRATING DIFFERENT BRANCHES INTO BASE BLOCK

Row	3×3 CONV	3×3 C-CDDC (DG)	3×3 C-CDDC (HV)	Metrics			
				mSA	Success	Precision	NPrecision
1	-	-	-	0.684	0.667	0.856	0.816
2	✓	-	-	0.695	0.677	0.865	0.821
3	-	✓	-	0.702	0.684	0.881	0.834
4	-	-	✓	0.708	0.690	0.885	0.839
5	-	✓	✓	**0.715**	**0.692**	**0.887**	**0.841**
6	✓	✓	✓	0.702	0.684	0.874	0.829

TABLE IV

PARAMETERS, FLOPS, AND PERFORMANCE FOR DIFFERENT ATTENTION MODULATION MODULES

Metrics	GCAM	PCAM-16	PCAM-4	Ours
Parameters (M)	2.16	2.17	**2.24**	2.19
FLOPs (G)	7.56	7.67	**7.97**	7.57
mSA	0.707	0.709	0.713	**0.724**
Success	0.689	0.691	0.694	**0.706**
Precision	0.886	0.881	0.898	**0.905**
NPrecision	0.840	0.837	0.843	**0.858**

different challenging scenarios. The combination of two directions C-CDDC paths yields the best performance which demonstrates their compatibility with each other. In addition, we observe that introducing normal 3 × 3 convolution in combination with the contrast-based branch results in a decline in model performance. We argue that this is due to the C-CDDC relying on sparse sampling points, while normal convolution samples densely, resulting in reduced contrast differences of targets in different directions.

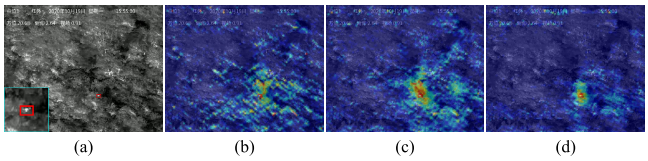


Fig. 12. Visualization of Grad-CAM [49] results before and after applying channel attention with compressed and uncompressed channel dimensions. (a) Input image, (b) PCAM-16, (c) PCAM-4, and (d) FFMM based on the PCDC module. The channel compression ratio consistently decreases from (b) to (d), leading to a progressively enhanced focus of the model on the target.

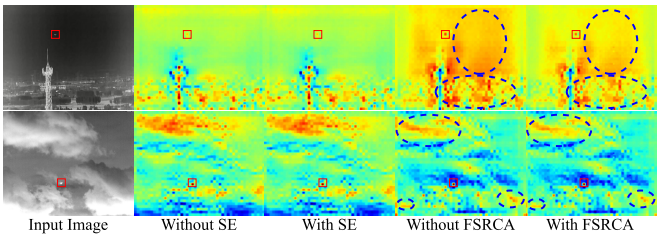


Fig. 13. Visualization of response maps before and after applying channel attention with compressed and uncompressed spatial dimensions. The first column represents the input images. The second and third columns illustrate the response maps before and after incorporating the SE attention mechanism. The fourth and fifth columns display the response maps before and after integrating the FSRCA.

3) Effectiveness of Integrating Pixel-Wise Attention Block: To demonstrate the effectiveness of FFMM, we conducted an ablation study on the different channel attention modules. Table IV shows the parameters, floating-point operations per second (FLOPs), mSA, success, precision, and NPrecision of different channel attention modulation modules. As illustrated in Table IV, GCAM has the worst performance. This shows that the simultaneous compression of spatial and channel dimensions significantly harms the infrared small UAV target information. The performance of PCAM can be improved as the channel compression ratio is reduced from 16 to 4. Compared with these attention mechanisms, our channel attention module simultaneously without spatial and channel compression delivers the best performance in terms of mSA, success, and precision. As shown in rows 1 and 3 of Table II, the mSA, success, precision, and NPrecision improve by 4.0%, 3.9%, 4.9%, and 4.2%, respectively. With our pixel-by-pixel channel attention mechanism, the target regions can be correctly strengthened while the background information can be highly suppressed, which leads to higher tracking performance. Additionally, the parameters and FLOPs of our methods are less than those of the PCAM-4. This is largely attributed to the efficient integration of pixel-wise cross-channel interaction based on depth-wise convolution.

Class activation maps of different channel attention modulation modules are shown in Fig. 12. Fig. 12(b) and (c) show the ability of the PCAM to suppress the background gets stronger when the compression ratio decreases. As shown in Fig. 12(d), our approach achieves remarkable efficacy in precisely focusing on IRUTs while effectively suppressing complex forest backgrounds, indicating our FSRCA module possesses better feature representation capability for small-scale UAV

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 05,2024 at 01:19:02 UTC from IEEE Xplore. Restrictions apply.

TABLE V
EFFECTIVENESS OF k ON THE PCDC MODULE

Row	Metrics			
	mSA	Success	Precision	NPrecision
k=3	**0.724**	**0.706**	**0.905**	**0.858**
k=5	0.722	0.703	0.903	0.856
k=7	0.716	0.698	0.891	0.845
k=9	0.720	0.703	0.896	0.852

TABLE VI
PERFORMANCE OF THE TRACKING METHOD IN TERMS OF NUMBER OF DYNAMIC TEMPLATES

Metrics	Number of dynamic templates				
	1	**2**	3	4	5
mSA	0.743	**0.748**	0.745	0.742	0.738
Success	0.725	**0.729**	0.726	0.723	0.719
Precision	0.931	**0.936**	0.932	0.930	0.925
NPrecision	0.880	**0.884**	0.882	0.879	0.874

targets and is able to adaptively select the distinctive features of IRUTs.

Fig. 13 visualizes the feature maps which is generated from the last block of stage 1 in the backbone network before and after applying FSRCA from Seqs. 3 to 6. SENet only uses one value to represent the importance of the channel and assigns the same importance to each pixel of the full-resolution feature map. Therefore, it is difficult to focus on the target and suppress the background at the same time. Columns 2 and 3 show the feature maps before and after applying SE attention to the feature maps. We can see that the overall feature information has not changed much, which suggests it is challenging to take into account all pixels in detail by using only one pixel to represent the information of the whole feature map. Columns 4 and 5 show the feature maps before and after applying FSRCA to the feature maps. From Seq. 6, it can be seen that the features of the bright cloud part are well suppressed. Compared to representing the overall feature map characteristics with a single value, the FSRCA can provide a better perception of targets and suppress backgrounds at a finer level.

Table V shows the impact of the number of channels for local interaction on performance. We can see that the performance peaks when $k = 3$, indicating using three local interaction channel numbers is sufficient to capture fine-grained small UAV target details effectively.

4) Effectiveness of Integrating Online Updating Module: From rows 1 and 4 of Table II, we can observe that the incorporation of CUFIM results in an enhancement of 2.1% in mSA, 2.0% in success, 3.0% in precision, and 2.1% in NPrecision which suggests only utilizing the initialized template can hardly meet the needs of long-term infrared UAV tracking task under complex background. By exploiting temporal contextual information, the model can accurately capture UAV targets with the perception of the changes in the target state and the surrounding background.

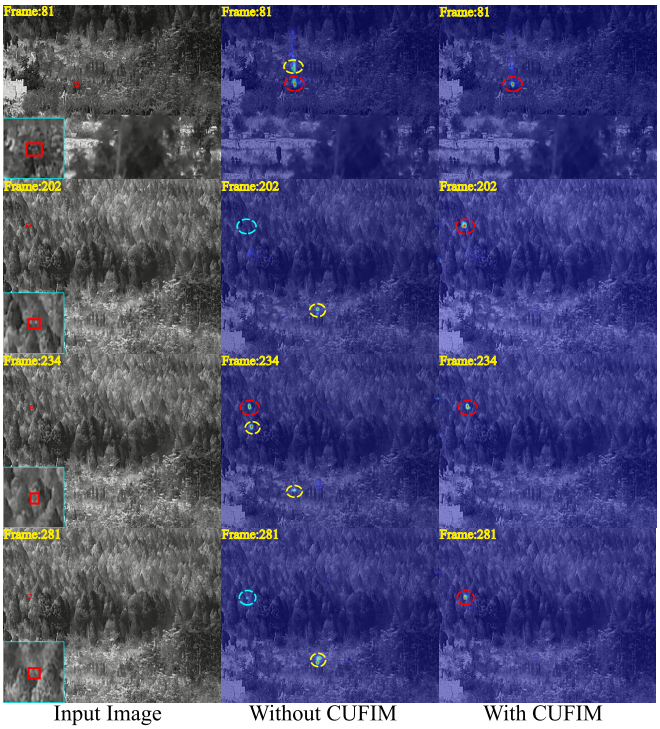


Fig. 14. Visualization of the confidence maps before and after adding our CUFIM. The red box indicates the tracking target. The red dotted box represents the true target, while the yellow and cyan dotted boxes represent the erroneous peak score predicted by the model and missed detection, respectively. Without CUFIM, the model is easily interfered by similar objects.

Table VI shows that using two extra templates brings the overall best performance. We argue that a single dynamic template may not provide enough information to adequately represent the state of the target. Furthermore, introducing more templates does not necessarily lead to better results when the number of templates exceeds two. We speculate that when too many templates are introduced, the model tends to combine the template that is most similar to the current target to be tracked, rather than fully utilizing the targets with different states.

Fig. 14 shows that the small UAV target flies in the forest background. Column 1 displays different frames in the same video sequence, and columns 2 and 3 visualize the confidence maps before and after applying CUFIM. As can be seen from column 2, without introducing an online updating module, the tracker easily drifts to the distractors causing peak confidence offset or present multiple peaks. The third and fourth confidence maps in column 2 exhibit multiple peaks, which implies that the baseline method is susceptible to being misled by distractors. Moreover, the second confidence map in column 2 demonstrates that the model is completely influenced by similar objects. The primary reason is that the baseline method lacks awareness of changes in the state of the IRUT and the changing background where it is located. From column 3, it is apparent that the high confidence score does not drift to any of the distractors, indicating that multitemplate information significantly helps the model to resist complex background interference.

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 05,2024 at 01:19:02 UTC from IEEE Xplore. Restrictions apply.

V. Conclusion

In this article, we propose a novel real-time IRUT long-term tracker (SiamCAP) for anti-UAV surveillance. First, we propose a new CMORB to effectively extract difference information and local contextual target features without introducing extra computational overhead. Second, we propose a novel FSRCA module without spatial and channel compression to pixel-wisely enhance IRUT features. Based on the FSRCA, we construct an FFMM to encode the low-level contextual information of IRUT details into high-level features in a bottom-up manner. Third, we propose a CUFIM that is capable of effectively capturing the interdependence between the multitemplate and the search frame while effectively exploiting the temporal contextual information. Our method consistently achieves the best accuracy-speed trade-off compared to the SOTA baseline tracking methods on real infrared UAV datasets and is potentially applicable to real-world anti-UAV tracking tasks.

We employ the confidence score to determine when to update the template. However, the distractor with a high score is also taken into consideration, which can potentially degrade the tracking performance. In the future, we will explore the design of an anti-distractor updating module which aims to improve the template update process by considering various factors beyond just the confidence score.

Acknowledgment

The authors would like to thank Zengyang Li and Ruixuan Huang for their help in performing experiments.

References

[1] L. Huang, X. Zhao, and K. Huang, "GlobalTrack: A simple and strong baseline for long-term tracking," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11037–11044.

[2] B. Yan et al., "Towards grand unification of object tracking," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2022, pp. 733–751.

[3] H. Fang, Z. Liao, X. Wang, Y. Chang, and L. Yan, "Differentiated attention guided network over hierarchical and aggregated features for intelligent UAV surveillance," *IEEE Trans. Ind. Informat.*, vol. 19, no. 9, pp. 9909–9920, Sep. 2023, doi: 10.1109/TII.2022.3232777.

[4] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional Siamese networks for object tracking," in *Proc. ECCV Workshops*, Oct. 2016, pp. 850–865.

[5] R. Tao, E. Gavves, and A. W. M. Smeulders, "Siamese instance search for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1420–1429, doi: 10.1109/CVPR.2016.158.

[6] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with Siamese region proposal network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8971–8980.

[7] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "SiamRPN++: Evolution of Siamese visual tracking with very deep networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4282–4291.

[8] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ATOM: Accurate tracking by overlap maximization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4660–4669.

[9] Z. Cao, C. Fu, J. Ye, B. Li, and Y. Li, "HiFT: Hierarchical feature transformer for aerial tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 15457–15466.

[10] B. Ye, H. Chang, B. Ma, S. Shan, and X. Chen, "Joint feature learning and relation modeling for tracking: A one-stream framework," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2022, pp. 341–357.

[11] S. Gao, C. Zhou, C. Ma, X. Wang, and J. Yuan, "AiATrack: Attention in attention for transformer visual tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 146–164.

[12] S. Gao, C. Zhou, and J. Zhang, "Generalized relation modeling for transformer tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 18686–18695.

[13] B. Yan, H. Zhao, D. Wang, H. Lu, and X. Yang, "'Skimming-perusal' tracking: A framework for real-time and robust long-term tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2385–2393.

[14] K. Dai, Y. Zhang, D. Wang, J. Li, H. Lu, and X. Yang, "High-performance long-term tracking with meta-updater," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6298–6307.

[15] P. Voigtlaender, J. Luiten, P. H. S. Torr, and B. Leibe, "Siam R-CNN: Visual tracking by re-detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6578–6588.

[16] B. Huang et al., "SiamSTA: Spatio-temporal attention based Siamese tracker for tracking UAVs," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1204–1212.

[17] J. Zhao, X. Zhang, and P. Zhang, "A unified approach for tracking UAVs in infrared," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1213–1222.

[18] N. Jiang et al., "Anti-UAV: A large-scale benchmark for vision-based UAV tracking," *IEEE Trans. Multimedia*, vol. 25, pp. 486–500, 2023.

[19] H. Wu, W. Li, W. Li, and G. Liu, "A real-time robust approach for tracking UAVs in infrared videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 4448–4455.

[20] H. Fang, X. Wang, Z. Liao, Y. Chang, and L. Yan, "A real-time anti-distractor infrared UAV tracker with channel feature refinement module," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, p. 1240.

[21] D. Guo, J. Wang, Y. Cui, Z. Wang, and S. Chen, "SiamCAR: Siamese fully convolutional classification and regression for visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6269–6277.

[22] X. Wang, J. Tang, B. Luo, Y. Wang, Y. Tian, and F. Wu, "Tracking by joint local and global search: A target-aware attention-based approach," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 11, pp. 6931–6945, Nov. 2022.

[23] X. Xu, J. Zhao, J. Wu, and F. Shen, "Switch and refine: A long-term tracking and segmentation framework," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 3, pp. 1291–1304, Mar. 2023.

[24] Y. Shan, X. Zhou, S. Liu, Y. Zhang, and K. Huang, "SiamFPN: A deep learning method for accurate and real-time maritime ship tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 315–325, May 2021.

[25] H. Bao, P. Shu, H. Zhang, and X. Liu, "Siamese-based twin attention network for visual tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 2, pp. 847–860, Feb. 2023.

[26] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Asymmetric contextual modulation for infrared small target detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Oct. 2021, pp. 950–959.

[27] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Attentional local contrast networks for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9813–9824, Nov. 2021.

[28] T. Zhang, L. Li, S. Cao, T. Pu, and Z. Peng, "Attention-guided pyramid context networks for detecting infrared small target under complex background," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 59, no. 4, pp. 1–13, Jul. 2023.

[29] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," 2018, *arXiv:1805.10180*.

[30] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[31] Q. Liu, X. Li, Z. He, N. Fan, D. Yuan, and H. Wang, "Learning deep multi-level similarity for thermal infrared object tracking," *IEEE Trans. Multimedia*, vol. 23, pp. 2114–2126, 2021.

[32] M. Hu et al., "Online convolutional reparameterization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 558–567.

[33] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Aug. 2022, pp. 11976–11986.

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 05,2024 at 01:19:02 UTC from IEEE Xplore. Restrictions apply.

[34] Z. Yu, Y. Qin, H. Zhao, X. Li, and G. Zhao, "Dual-cross central difference network for face anti-spoofing," in *Proc. 13th Int. Joint Conf. Artif. Intell.*, 2021, pp. 1281–1287, doi: 10.24963/ijcai.2021/177.

[35] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.

[36] B. Yan, X. Zhang, D. Wang, H. Lu, and X. Yang, "Alpha-refine: Boosting tracking performance by precise bounding box estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5289–5298.

[37] X. Chen, B. Yan, J. Zhu, H. Lu, X. Ruan, and D. Wang, "High-performance transformer tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 8507–8523, Jul. 2023.

[38] Z. Zhang, H. Peng, J. Fu, B. Li, and W. Hu, "Ocean: Object-aware anchor-free tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 771–787.

[39] G. Jocher et al. (2021). *YOLOv5*. [Online]. Available: https://github.com/ultralytics/yolov5

[40] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[41] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 658–666.

[42] (2023). *Anti-UAV Challenge Dataset*. Accessed: Aug. 23, 2023. [Online]. Available: https://anti-uav.github.io/dataset/

[43] J. Zhao et al., "The 2nd anti-UAV workshop challenge: Methods and results," 2021, *arXiv:2108.09909*.

[44] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2411–2418.

[45] M. Müller, A. Bibi, S. Giancola, S. Alsubaihi, and B. Ghanem, "TrackingNet: A large-scale dataset and benchmark for object tracking in the wild," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Jun. 2018, pp. 300–317.

[46] R. Kou, C. Wang, Y. Yu, Z. Peng, F. Huang, and Q. Fu, "Infrared small target tracking algorithm via segmentation network and multistrategy fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023.

[47] Q. Liu et al., "LSOTB-TIR: A large-scale high-diversity thermal infrared object tracking benchmark," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 3847–3856.

[48] H. Fang, L. Ding, L. Wang, Y. Chang, L. Yan, and J. Han, "Infrared small UAV target detection based on depthwise separable residual dense network and multiscale feature fusion," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–20, 2022, doi: 10.1109/TIM.2022.3198490.

[49] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.



Houzhang Fang (Member, IEEE) received the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2014.

He is currently an Associate Professor with the School of Computer Science and Technology, Xidian University, Xi'an, China. His research interests include target recognition, target tracking, and image restoration.



Chenxing Wu received the bachelor's degree from the School of Computer Science and Technology, Xidian University, Xi'an, China, in 2022, where he is currently pursuing the master's degree.

His research interests include object tracking, object detection, and anti-UAV technologies.



Xiaolin Wang received the bachelor's and master's degrees from the School of Computer Science and Technology, Xidian University, Xi'an, China, in 2020 and 2023, respectively, where he is currently pursuing the Ph.D. degree.

His research interests include object tracking, object detection, infrared image processing, and anti-UAV technologies.



Fan Zhou received the master's degree from Xidian University, Xi'an, China, in 2006.

He is currently a Senior Engineer with Xi'an Modern Control Technology Research Institute, Xi'an. His research interests include target recognition and guidance system effectiveness evaluation.



Yi Chang (Member, IEEE) received the B.S. degree in automation from the University of Electronic Science and Technology of China, Chengdu, China, in 2011, and the M.S. degree in pattern recognition and intelligent systems and the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2014 and 2019, respectively.

He is currently an Associate Professor with the School of Artificial Intelligence and Automation, HUST. His research interests include image processing, computer vision, and machine learning.



Luxin Yan (Member, IEEE) received the B.S. degree in electronic communication engineering and the Ph.D. degree in pattern recognition and intelligence system from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2001 and 2007, respectively.

He is currently a Professor with the School of Artificial Intelligence and Automation, HUST. His research interests include multispectral image processing, pattern recognition, and real-time embedded systems.

Authorized licensed use limited to: XIDIAN UNIVERSITY. Downloaded on August 05,2024 at 01:19:02 UTC from IEEE Xplore. Restrictions apply.